Evaluation Matters—He Take Tō Te Aromatawai 1: 2015
© New Zealand Council for Educational Research 2015
http://dx.doi.org/10.18296/em.0007
http://www.nzcer.org.nz/nzcerpress/evaluation-matters

Evaluative reasoning in public-sector evaluation in Aotearoa New Zealand: How are we doing?

Heather Nunns, Robin Peace, and Karen Witten

This article reports the results of a meta-evaluation of 30 publicly accessible evaluation reports written or commissioned by 20 New Zealand public-sector agencies during the period 2010–2013 to understand how evaluative reasoning is being practised in Aotearoa New Zealand. The reports were examined to find evidence of five key elements of evaluative reasoning, namely, evaluative objectives or questions, criteria or other comparator(s), defined standards, a warranted argument, and an evaluative conclusion or judgement. Only eight of the evaluation reports had evidence of all five elements. While the focus of the meta-evaluation was on the presence of the five elements (not their quality) and the report sample is not representative, the study provides an interesting snapshot of evaluative practice in Aotearoa New Zealand. The findings suggest there may be scope to improve evaluative reasoning practice.

Introduction

Professional evaluators are concerned about the quality of the evaluations they undertake, ensuring the appropriateness of the evaluation design and sample, the quality of evidence collected, and the correct application of the methods used. This article argues that such dimensions of quality are necessary but not sufficient, and that the crux of evaluation quality is sound evaluative reasoning. Drawing on the work of Scriven (1991), Fournier (1995) defines evaluative reasoning as "the systematic means for arriving at evaluative conclusions, the principles that support inferences drawn by evaluators" (p. 1). Without sound evaluative reasoning, all other efforts aimed at producing quality evaluations are compromised. Public-sector decision makers need robust and defensible answers to evaluative questions. Such answers are underpinned by sound evaluative reasoning.

This article describes evaluative research to understand how evaluative reasoning is being practised in public-sector evaluation in Aotearoa New Zealand. It is based on a meta-evaluation of 30 evaluation reports in the public domain which were conducted or commissioned by 20 central government agencies during the period 2010–2013. The meta-evaluation, part of an ongoing study examining how evaluative reasoning is understood and practised in New Zealand public-sector evaluation, is framed by five evaluative criteria derived through the research underpinning this study.

Context

Evaluation is sometimes viewed as a professional practice, rather than as a discipline with a theoretical base (Donaldson & Lipsey, 2006). This article is premised on evaluation as a discipline, based on theory derived from western philosophy. Up until the mid-20th century, the rules of formal logic made it logically impossible to reason from a factual premise to an evaluative claim (Scriven, 2013). Developments in informal logic such as those articulated by Hare (1967), Rescher (1969), and Taylor (1961) meant reasoning about values became logically possible, as summarised in the left hand column in Table 1.

These elements, articulated by Scriven (1980, 1991, 1995) became known in the evaluation literature as the general logic of evaluation (summarised in the right hand column of Table 1).

Table 1. Informal logic: How to reason evaluatively

Hare (1967), Rescher (1969), Taylor (1961)	Scriven's general logic of evaluation (1980, 1991, 1995)
1. Identify the object (X) and the value to be applied to the object 2. Identify the "class of comparison" to which X belongs (Z) 3. Identify norms for Z	1. Establish criteria of merit for the evaluand
4. Develop a set of operational statements describing levels of performance for each of the norms of Z	2. Construct standards for the criteria
5. Determine the characteristic(s) of X (the "good making characteristics")	3. Measure performance of the evaluand against the criteria 4. Synthesise and integrate data into a judgement of merit or worth
6. Compare X's characteristics with the operational statements above to come to an evaluative conclusion	
7. Justify the norms used	

The elements shown in Table 1 have been further explicated by evaluation theorists resulting in a body of knowledge about what is required to reason from a value to an evaluative conclusion that is valid and robust, referred to as evaluative reasoning (House & Howe, 1999, p. xvi). The centrality of evaluative reasoning to the practice of evaluation is emphasised in this literature. According to House (1980), evaluative reasoning is "the substance of evaluation" (p. 5), a view reinforced by Patton (2012) who states: "valuing is fundamentally about reasoning and critical thinking. Evaluation as a field has become methodologically manic-obsessive. Too many of us, and those who commission us, think that it's all about methods. It's not. It's about reasoning" (p. 105).

The role of evaluation in the development and assessment of public policy is well documented (Chelimsky, 2012; Grob, 2003). Evaluation provides information about what works, for whom, and why, as well as determining whether the desired outcomes and impacts of public policy are being achieved. For evaluation to be viewed as a credible contributor to public policy, evaluative conclusions need to be robust. Further, because evaluative judgements are "consequential" (Greene, 2011, p. 90), they need to be defensible.

Conceptual framework

A conceptual framework was developed from the literature consisting of five interconnected elements: evaluative objectives or questions, criteria or other comparator(s), defined standards, warranted argument, and an evaluative conclusion or judgement. These five elements work together to build a coherent case to support an evaluative claim from which an evaluative conclusion can be drawn that is legitimate and defensible (Fournier, 1995; Fournier & Smith, 1993) as illustrated in Figure 1. The interconnected nature of the elements is as follows. The evaluation objectives or questions provide the purpose and focus for the evaluation. They also determine the choice of the criteria (or other comparator) against which the evaluand is to be examined. Standards are required for the criteria (or other comparator) to identify and describe levels of performance. A warranted argument is required to support and strengthen the evaluative claim(s) about the performance of the evaluand (i.e., the evidence) in relation to the criteria and standards. The warranted argument sets out the case from which can be drawn an evaluative conclusion/judgement. Each element is now described more fully.

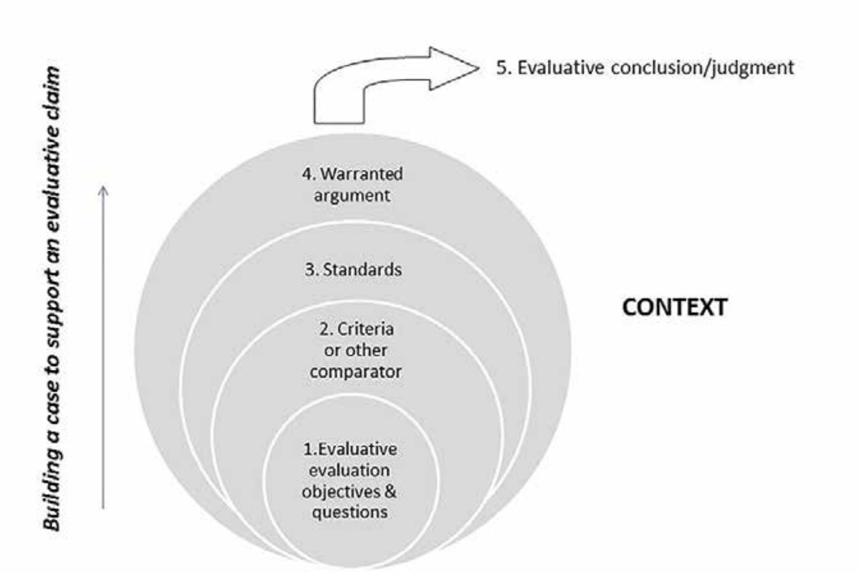
Building a case to support an evaluative claim
5. Evaluative conclusion/judgment
4. Warranted argument
3. Standards
2. Criteria or other comparator
1.Evaluative evaluation objectives & questions
CONTEXT

Figure 1. Building a case to support an evaluative claim: Elements of evaluative reasoning

Evaluation objectives and questions: The evaluation objectives and questions focus the inquiry by providing the foundations for how the evaluation is designed, conducted, and reported. They determine the values to be examined, the criteria and standards to be selected, the data to be collected, and the nature of the argument required to support the evaluative conclusion/judgement. There are two types of evaluation objectives and questions: evaluative (those containing a value word), and non-evaluative (those containing descriptive or explanatory language) (Davidson, 2005). This study makes a distinction between evaluative research (research that answers evaluative questions) and non-evaluative research (research that answers other types of questions, such as descriptive or explanatory questions). An evaluation that contains mostly non-evaluative objectives is likely to be the latter.

Criteria or other comparator(s): To evaluate is to compare: "The fundamental idea of conceptualising quality is through comparison, direct or even vaporously indirect" (Stake & Schwandt, 2006, p. 412). Criteria provide the most explicit approach for such comparison. Criteria are the "aspects, qualities or dimensions that distinguish a more meritorious or valuable evaluand from one that is less meritorious or valuable" (Davidson, 2005, p. 91). Criteria provide the grounds on which the evaluator reasons towards an evaluative conclusion/judgement (Fournier, 1995; Valovirta, 2002). The critical contribution of criteria to the evaluative judgement is described in the literature. Fournier (1995) observes "criteria can make or break an evaluation because they … directly affect the validity of claims" (p. 19), while Valovirta (2002) notes "the grounds (criteria) on which evaluative judgments have been made form the basis of one of the most common forms of debate about an evaluation report" (p. 63). The validity of criteria (and therefore of the evaluative conclusion or judgement) is strengthened if the criteria are justified, for example by reference to their source (Hurteau, Houle & Mongiat, 2009). Stake and Schwandt (2006) exhort evaluators to be rigorous in their approach to identifying criteria: "Majority opinion (of stakeholders) should not be considered sufficient … standards of quality generated by representative groups and quotations from learned papers are but starting points" (p. 412).

While the use of criteria is encouraged by some evaluation theorists (for example, Davidson, 2005), Stake and Schwandt (2006,) describe ways of making a comparison that are less likely to promote what they term "criterial thinking" or "quality-as-measured" (p. 407). These approaches are based on practical experience of an evaluand through perceptual and experiential knowledge. They refer to such approaches as "quality-as-experienced" (Stake & Schwandt, 2006, p. 407), for example, the assessment of wine by the expert viticulturist, or the valuation of fine art by the professional appraiser.

Standards: While criteria define quality, they require accompanying standards to explicate how quality is discerned in relation to *better* quality and *poorer* quality. Standards act as benchmarks (Arens, 2005, p. 18) against which the evaluand can be compared and ranked. Standards may be expressed quantitatively (for example, by a number, grade, or rank), or qualitatively (such as from "inadequate" to "excellent") (Davidson, 2005, pp. 137, 142).

Warranted argument: Argument is an essential element of evaluative reasoning because it articulates the inference that links evidence to an evaluative claim (Fournier & Smith, 1993). According to Fournier and Smith (1993), "building a justifiable argument is the crux of evaluation practice" (p. 316). Argument assumes greater importance in evaluation than in other forms of systematic inquiry because of the type of inference used, namely, probative inference. This type of inference leads to a conclusion that is not certain (in a deductive sense), but rather is an "all things considered" judgement (House, 1995, p. 40) or a "prima facie conclusion" (Scriven, 1991, p. 277). A well-constructed and supported argument builds the plausibility of the claim (Booth, Colomb & Williams, 2008, p. 112). The importance of argument in arriving at an evaluative conclusion/judgement is stressed by Schwandt (2008): "My concern is that in the press to master methods of generating data, we ignore the idea of developing a warranted argument—a clear chain of reasoning that connects the grounds, reasons or evidence to an evaluative conclusion" (p. 146).

A warrant is the "because" part of an argument. It legitimates the inference from the claim and evidence to the conclusion by referring to an appropriate authority. Warrants vary across disciplines and are context-dependent (Smith, 1995). For example, lawyers use legal precedence as a warrant, physical scientists rely on the laws of nature (such as the law of gravity), and artists rely on expert opinion (Toulmin, Rieke & Janik, 1979). The inclusion of a warrant that is both reliable and relevant to the claim adds further weight by establishing legitimacy through appealing to an authority or general principle (Booth et al., 2008, p.116, p.157). Greene (2011) is unequivocal about the need for evaluators to be diligent in warranting their evaluative claims and argument: "Worrying about warrant is a core evaluator responsibility. It is because our inferences are consequential that we must have confidence that they are warranted" (p. 90).

Evaluative conclusion or judgement: This is the intended endpoint or destination of the reasoning process. Stake and Schwandt (2006) describe judgement-making as being fundamental to the evaluation profession: "Making judgments of quality constitutes a core professional responsibility of evaluators" (p. 416). Despite this, House and Howe (1999) note that "there are no clear professional rules" (p. 30) available to the profession about how to do so. According to these authors, judgement-making cannot be reduced to a set of standardised procedures. Rather, judgement-making requires the evaluator "to take relevant multiple criteria and interests, and combine them into all-things-considered judgements in which everything is consolidated and related" (House & Howe, 1999, p. 29). For an evaluative judgement to be legitimate and defensible, there must be coherent and transparent connections across the evaluation objectives or questions, criteria (or other comparator(s)) and standards, claims, argument, and judgement (Fournier, 1995; Fournier & Smith, 1993). Finally, the contingent nature of evaluative judgements must be stressed. Schwandt and Stake (2006) remind evaluators that their conclusions or judgements are "perspectival, temporal and conditional" (p. 412).

In basing the conceptual framework on these five elements, it is not intended to simplify or reduce evaluative reasoning to an easy-to-master technique. The complexity that is evaluative reasoning is acknowledged. Authors describe evaluative reasoning as involving careful listening (Abma, 2006), perceptive consideration of stakeholder perspectives (Stake, 2004), reflexivity about personal values and their impact on the deliberative process (Greene, 2011), critical thinking (Schwandt, 2001), sensitivity to bias (Denzin & Lincoln, 2011), and astute judgement-making (Scriven, 1994). The five elements provide a framework around which such complexity is built.

Figure 1 shows the elements of evaluative reasoning occurring in a context. The influence of context on evaluative reasoning must not be underestimated: "Valuing must be understood as contextually embedded and dependent" (Patton, 2012, p. 98). Context is defined as referring to "The setting within which the evaluand … and thus the evaluation are situated. Context is the site, location, environment or milieu for a given evaluand" (Greene, 2005, p. 83). To evaluate is to confront context. The programmes, policies and strategies we evaluate are not discrete, detached constructs but arise from and exist within a context: "Evaluands are social, political and moral constructions that embody the different (and often conflicting) interests and values of stakeholders (Schwandt, 1997, p. 26). Most significantly, evaluations commissioned, or funded, or both, by public-sector agencies will be determined by the priorities and interests of the government of the day. The political context determines what type of evidence is valued and therefore what evaluation methods are regarded as valid and trustworthy. At a practice level, context determines the choice of criteria, how they are developed and by whom (Henry, 2002), the nature and extent of argument, the warrants selected (Smith, 1995), and the complexity of the evaluative conclusion/judgement required (Julnes, 2012).

Method

Meta-evaluation

A meta-evaluation approach was chosen to examine a sample of 30 public-sector evaluation reports. This entailed identifying criteria against which each report could be assessed. The five criteria are derived from the conceptual framework as follows: (i) evaluation objectives that are evaluative; (ii) criteria or some other comparator against which the evaluand is examined; (iii) standards that are defined; (iv) a warranted argument linking evidence and claims; and (v) a conclusion or judgement that is evaluative. We have chosen to use the term *elements* to describe these, so as to avoid confusion with the term criteria as used in the conceptual framework. The aim was not to examine the quality of these elements in collectively building an evaluative case, but rather to find evidence of their presence. Given our interest in identifying a range of practice, we decided to look for evidence of evaluative reasoning in a large sample of reports rather than examining a smaller number in greater detail.

As the research aim was to find evidence of the five elements in the reports, the standard is whether there is evidence of the element in the report. While this assessment was straightforward for elements (i), (ii), (iii), and (v), it was less straightforward for warranted argument (element iv). The definition of warranted argument used in this study is based on that of Booth, Colomb and Williams (2008, p. 109), who describe a research argument as consisting of five components: (1) a claim; (2) reasons that support the claim; (3) evidence that supports the reasons; (4) an acknowledgment of and a response to alternatives/complications/objections; and (5) a principle which makes the reasons relevant to the claim, referred to as *the warrant*. The argument in each report was examined to determine whether these components were addressed.

Given the importance of context in evaluative inquiry, the reports were examined for contextual information to inform understanding of the factors that influenced the design and conduct of the evaluation. Such information included the evaluation purpose, the audiences for and intended uses of the evaluation findings, information about the evaluand and context, the methods used, and limitations of the evaluation. The limitations information in the report was also examined to understand the contextual factors that may have impacted on the evaluation and its valuing approach. This information about individual reports was recorded in tabulated form to provide insight about the elements and their application.

Sampling approach

The websites of 52 central government agencies were searched. The search was restricted to reports dated 2010–2013 to ensure their currency. Evaluation reports were available on the websites of 22 agencies, while the websites of 30 agencies either had summaries of reports, reports dated pre-2010, or contained no reports. From the 53 evaluation reports collected, 30 reports were chosen based on the following criteria:

i. The reports were commissioned, or funded, or both, by New Zealand central government agencies.
ii. The reports were written by New Zealand authors.
iii. No author appears twice in the sample.
iv. The report is a complete evaluation report rather than a summary.
v. The sample includes a range of evaluand types (for example, policy, programme, media campaign, strategy) and evaluation approaches (for example, development, developmental, economic, implementation, outcomes, impact).
vi. There are no more than two reports per agency.

Of the 30 reports chosen, 11 reports have authors who were employed by the agency, referred to as internal authors. It was assumed the authors of the seven reports with anonymous authors were agency employees. Nineteen reports have authors who worked outside the agency, referred to as external authors. There is a range of evaluands in the sample. Social and educational programmes were the most common (17 reports), policies (4 reports), interventions (4 reports), an aid strategy (1 report), a media strategy (1 report), governance arrangements (1 report), a road construction project (1 report), and research use (1 report). The sample includes two health-impact assessments and five economic evaluations, two of which are cost–benefit analyses (CBA). Significant differences were found between the CBA studies and the three other economic evaluations. Specifically, the distinct features of the CBA method do not align with the five elements in the conceptual framework in the same way as do the two value-for-money and the one cost-effectiveness analyses. Owing to space constraints, the results pertaining to the two CBA studies have been excluded from this article, thereby reducing the sample to 28.

An evaluation orientation (Chelimsky, 1997) was assigned to each report based on its evaluation purpose statement, namely, accountability (the measurement of results or efficiency), development (the provision of evaluative help to strengthen institutions), knowledge (the acquisition of a more profound understanding in some specific area or field), and management (for oversight, or improvement, or both). The authors of the reports work in a range of professions, including civil engineering, economics, health, and management. They include education consultants, psychologists, health professionals, academics, and evaluation practitioners.

Results

This section begins by describing the evaluation contexts for the evaluations in the sample. The findings for the 28 reports are then summarised for each of the five elements of evaluative reasoning.

Evaluation contexts

The five elements of evaluative reasoning potentially provide a coherent framework to guide the evaluator's work. However, as any experienced practitioner will confirm, applying these elements in a practical context can be less than straightforward. This section illustrates the diversity of contexts in which evaluation takes place and, therefore, the range of influences and constraints on the evaluator's work which in turn may influence evaluative reasoning practices.

Seventeen evaluations were conducted in community settings in New Zealand: with Māori communities; with people living in temporary accommodation whose homes had been destroyed in the Canterbury earthquakes; with vulnerable parents in their homes; in marae-based courts; workplaces; therapeutic communities; and schools. Two evaluations of New Zealand-funded aid projects were conducted in community settings in Pacific Island countries. Such community-based settings often involve challenges for the evaluator, such as issues of respondent accessibility, time constraints, and resource availability. Compromises and trade-offs may have to be made which may impact on the evaluation, and therefore on evaluative reasoning practice. Such constraints are illustrative of what Patton (2012) refers to as "the contextual pragmatics of valuing" (p. 97), requiring flexibility and adaptability on the evaluator's part to ensure the optimal quality of the evaluative reasoning despite the constraints. In contrast to the 17 community-based evaluations, six evaluations were primarily desk-based using either only secondary data (four reports) plus minimal qualitative data (two reports). The remaining five evaluations were undertaken in organisational settings.

Following Chelimsky's (1997) evaluation orientations, the majority (17) of the 28 reports have a management orientation, 10 reports have an accountability orientation, one has a development purpose, while no reports have a knowledge orientation. The emphasis on instrumental purposes is not surprising given the public sector's focus on efficiency and effectiveness of public expenditure.

Accounts of the limitations of the evaluation were examined to understand how contextual and other constraints may have impacted on the evaluators' work. Given that it is usual research practice to identify the limitations associated with an inquiry, it was unexpected to find that half of the 28 reports contained no information about the limitations associated with the evaluation. The absence of information about limitations restricted understanding of the contextual and other factors that may have influenced the design and conduct of the evaluation.

Element one: Evaluation objectives and questions

The reports were examined for their evaluation objectives. Reports that do not contain objectives were examined for their key evaluation questions (KEQs) or evaluation questions. Of the 28 reports, 24 contain one or more evaluation objectives, KEQs or questions, of which seven reports contain only *evaluative* objectives, KEQs or questions, seven reports contain only *non-evaluative* objectives/KEQs/questions, while 10 reports have a *combination* of evaluative and non-evaluative objectives/KEQs/questions. Four reports have no objectives/ KEQs/questions. Table 2 provides examples of evaluation objectives/KEQs/questions from two evaluation reports in the sample—the first report has all non-evaluative objectives, and the second report has two evaluative KEQs.

Table 2. Evaluation objectives and questions from two reports

Report 1: A report with evaluation objectives that are all non-evaluative	1. To identify how the changes are working for councils. 2. Whether the changes have addressed the challenges and issues posed by the previous governance model. 3. To identify potential areas of support,where (name of agency) can assist councils and institutions to achieve the outcomes intended.
Report 2: A report with of KEQs that are evaluative	1. To what extent, and in what ways was the investment in (name of initiative) successful in enhancing the capacity and capability of (name of group)? 2. Was the investment delivered efficiently and effectively so as to contribute to successful impacts?

The 17 reports with some or all evaluation objectives/KEQs/questions that are evaluative provide a foundation for evaluative reasoning. As might be expected in public-sector evaluation, many of the values expressed in the evaluation objectives and questions are associated with concerns such as effectiveness, efficiency, relevance, sustainability, value for money, and cost-effectiveness.

Element two: Criteria or other comparator(s)

Values are defined in the form of criteria in 11 of the reports, in either a rubric format (six reports), descriptive textual definitions (three reports), or expressed as indicators (two reports). The reports were examined to identify how the criteria were identified (Table 3). Davidson (2005) notes that an individual criterion can be weighted according to importance or some other aspect. The criteria are given equal weight in all 11 evaluations which use criteria.

Table 3. Source of criteria

Source of Criteria	Reports (n)
Evaluation commissioner, or stakeholders, or both	5
Relevant legislation, literature , policy, and programme documentation	3
Existing criteria, e.g. child maltreatment prevention criteria	2
Criteria were developed by the authors who are subject experts about the evaluand, or because no relevant literature or studies were found	2

As described above, authors may prefer to use a comparator other than criteria. Three reports (all of which are evaluations of therapeutic and capability building interventions aimed at individuals) include relevant academic literature. In all three cases the literature is used to provide a broad context about the topic within which the findings about the evaluand are presented. Used in this way the literature functions as an indirect comparator, rather than specific findings being compared directly with relevant topics in the literature.

Of the 11 reports that lack evaluation objectives or questions, or which include non-evaluative objectives or questions, nine reports refer to one or more value terms in the body of the report, or in an evaluative conclusion/judgement. Such value terms are not defined. An example of one of these reports is as follows. The objective of an evaluation of a training programme is "to examine the extent to which the intent of the (name of programme) was met and identify what went well and what could be improved going forward". (There are no evaluation questions.) There is no definition or description of what the programme working well would look like. The report consists of a findings section identifying four aspects of the programme that have been successful and three aspects requiring improvement. The evaluative conclusion identifies the successful elements of the programme: "The evaluation found the following elements of the initiative were successful: community leadership, scholarship model, academic support, community pastoral care." The authors make statements about what went well and provide a judgement about success without these value terms being defined.

Element three: Standards

Of the eleven reports with criteria, six reports include standards of performance that are defined. In three reports, the definitions of standards are tailored to the evaluand, while three reports use standards based on generic definitions of performance. The remaining five reports include references to standards of performance, but the standards are not defined.

Element four: Warranted argument

Seventeen of the 28 reports contain an argument. That is, the author interprets the evidence to produce one or more claims that are supported by reasons and evidence. In contrast, 11 of the 28 reports either do not contain an argument (8 reports) or have text that is ambiguous. That is, it is not clear whether the text is evidence or argument (three reports). These 11 reports are described in more detail below.

It was assumed that the reports would follow the traditional structure of research reports, namely, a section presenting the evidence (in New Zealand this is usually referred to as the *findings* section), followed by a section interpreting the evidence in the form of claims and argument (this is usually referred to as the *discussion* section). This structure clearly delineates evidence from the evaluator's interpretation of it. Of the 17 reports with an argument, in 10 reports the argument is located in a separate section to the findings, while in seven reports the presentation of the evidence is combined with the argument. In at least half of these seven reports, there are places where it is difficult to differentiate between evidence and the authors' interpretation of the evidence. This has the effect of weakening the argument.

The authors of the eight reports without an argument summarise the evidence. This is followed by a short section, usually headed *conclusion*, which contains the authors' claims about the evidence. This section may finish with an evaluative conclusion/judgement. Some of these reports give the impression of the author as a narrator who reports the views of different stakeholders as evidence. The author then changes hat and becomes an evaluator, issuing an evaluative claim. There is a lack of explicit interpretation of, and argument about the evidence. As a result the inferential leap between evidence and claim is left to the reader to work out.

Thirteen of the 17 reports that contain an argument use one or more warrants in an explicit or implicit manner, as summarised in Table 4.

Table 4. Types and frequency of warrants used

Type of warrant	Reports (n)	Example
Literature or other relevant information	6	A value for money evaluation of a health-related intervention compares New Zealand's experience to research about four overseas jurisdictions.
Cultural warrant	4	Three evaluations of initiatives involving Māori are based on kaupapa Māori principles and are authored by Māori evaluators.
Methodological warrant	4	The evaluators held workshops on the findings with stakeholders to validate the data analysis
Expert warrant	3	The authors of an evaluation of an early childhood parenting intervention involved a child health expert in the data analysis.
Authority warrant	2	Education Review Office reports. The ERO has a statutory role as the agency responsible for evaluating the pre-compulsory and compulsory education sectors.

Element five: Evaluative conclusion/judgement

Twenty-four reports contain evaluative conclusions/judgements of one of three types, as summarised in Table 5. Of these 24 reports, 11 can be considered as having unsound evaluative conclusions/judgements in that they use value terms that are referred to elsewhere in the report but are not defined, or use value terms that are not referred to anywhere else in the report.

Table 5. Types of evaluative conclusion/judgement

Type of evaluative conclusion/judgement	**Reports (n)**
Based on criteria or other comparator(s)	13*
Based on value terms that are referred to in the report but are not defined	6
Based on value terms that are not referred to anywhere else in the report	5

** Of these reports, eight report by individual criteria/comparator; and five report by individual criteria/comparator which are also synthesised into an overall qualitative judgment. (No reports explain how the individual assessments were synthesised).*

Overview of results

The results of the meta-evaluation show that eight of the 28 reports have evidence of all five elements (Table 6). All but one of these evaluations was written by external authors. Eleven reports demonstrate three or four of the elements. The most common omission is that value terms referred to in the report are not defined, for example, by criteria, indicators, or in a descriptive textual definition (seven reports). There is no significant difference in authorship of these 11 reports—six were authored by external authors and five by internal authors. The final group is made up of nine reports which lack three or more of the five elements. Surprisingly, three of these reports end with a conclusion/judgement that uses evaluative language despite an absence of most or all of the preceding elements.

Table 6. Results by author type

	Reports by number of elements of evaluative reasoning		
	five elements	three–four elements	two or fewer elements
Internal authors	1	5	5
External authors	7	6	4
Total	8	11	9

The three groups of reports shown in Table 6 were analysed to ascertain whether any patterns were discernible, for example, by evaluand type, evaluation orientation, or approach. No trends were apparent among these dimensions.

Discussion

While the meta-evaluation is not representative of public-sector evaluation practice in Aotearoa New Zealand, it provides a *snapshot* of evaluation practice and as such offers insights for further consideration and investigation. The findings suggest there may be variable practice in evaluative reasoning among authors of public-sector evaluations and this section offers possible explanations for the observed variability.

The first explanation concerns the authors in the evaluation sample. As noted above, the authors of the 28 reports are working in a range of professional areas. It is surmised that some authors may not identify professionally as evaluators and therefore may not be aware of evaluation theory and its implications for evaluation practice.

A different explanation is offered in respect of the authors in the sample who identify professionally as evaluators. We speculate that in Aotearoa New Zealand, as others have observed elsewhere, the evaluation community has been preoccupied with practice. We suggest this has distracted evaluators from becoming engaged with evaluation theory, and therefore with evaluative reasoning. In his presidential address to the 1998 conference of the American Evaluation Association, William Shadish (1998) stated: “Perhaps because evaluation is primarily a practice-driven field, evaluation theory is not a topic that causes evaluators’ hearts to flutter” (p. 1). The same could be said about the Aotearoa New Zealand situation—the evaluation community has developed through doing evaluation rather than learning about the theory underpinning it.

The emphasis given by successive governments to the role of evidence in public-sector decision making may have contributed to the evaluation community's preoccupation with practice-based topics, rather than evaluation theory. The terms "evidence-based practice" (Nutley, Davies & Walter & 2003) and "evidence-informed policy making" (Gluckman, 2013) describe this discourse. New Zealand's focus on evidence has followed that of other countries such as the United Kingdom where the Labour Government released a white paper titled *Modernising Government* (Prime Minister and Minister for the Cabinet Office, 1999) endorsing evidence-based policy making, and Australia where the Australian Productivity Commission has focussed on evidence-based policy making (Scobie, 2009).

We suggest that this focus on practice and practice-based topics such as evidence have preoccupied the evaluation community for some 10 years. This observation does not undervalue their importance in evaluative enquiry. However this focus has had the effect of distracting evaluators from engaging with the theory that underpins evaluation, and therefore from evaluative reasoning.

Limitations

There are a number of limitations associated with this meta-evaluation. First, the sample is not, nor does it claim to be, representative of public-sector evaluation reports. Only evaluation reports in the public domain were examined. There is no requirement in New Zealand for public-sector agencies to make evaluation reports and other official documents publicly available, other than via a formal request made under the Official Information Act 1982. As noted above, of the 52 agency websites examined, only 22 websites contained evaluation reports (not summaries) dated 2010–2013. The websites of the remaining 30 agencies either had summaries of evaluation reports, reports dated pre-2010, or contained no reports. No reports were available on the websites of some large agencies such as the New Zealand Qualifications Authority which evaluates the performance of tertiary institutions other than universities. Secondly, while there are many programme evaluation reports posted on websites, there are noticeably fewer policy evaluation reports. The reports of some large-scale, national policy initiatives posted on agency websites (such as Working for Families and KiwiSaver) are summary reports of high-level findings written for a general audience. Such reports were excluded from the sample because they lack sufficient detail about the evaluation approach required for this study. Lastly, the study focuses only on central government agencies as there was an insufficient number available on local government agency websites.

A further limitation is associated with a desk-based examination. There are a range of influences that determine the design, conduct, and reporting of an evaluation about which a desk-based study lacks information. As noted above, political influences are inherent and significant in the public-sector context. What is presented in an evaluation report represents the requirements of the commissioner (whether the report has been produced internally or externally). An evaluation may have been poorly scoped by the commissioning agency, leaving the external evaluator to do the best they can with a poorly considered brief. Other possible scenarios include a commissioner asking for a report that presents only the high-level findings, with lesser value being placed on a robust argument supporting such findings. An evaluation may have had a limited budget, resulting in an argument that is unable to be supported by expert or literature-based warrants. These and other important contextual factors are not visible in an evaluation report unless they are identified and discussed as limitations.

Conclusion

Of the five elements of evaluative reasoning in the reports examined, warranted argument is the element which appears to be most neglected. Eleven of the 28 reports either do not contain an argument or contain text that is ambiguous, that is, it is not clear whether the text refers to evidence or is the authors' argument. A further seven reports combine evidence and the authors' interpretation of the evidence (argument). At least half of these seven reports contain text where it is difficult to differentiate evidence from argument. Consequently, around half of the 28 reports lack an argument, or they have text which is ambiguous. This is a significant shortcoming that undermines the defensibility of the evaluative conclusion/judgement, exposing the evaluation to criticism about its validity and quality.

Further research is needed to test whether the results in this study are confirmed. Such research could include interviewing the reports' authors to understand whether (and how) factors such as context, funding, time constraints, and commissioner requirements influenced the evaluative reasoning underpinning the report. An extended desk-based study based on evaluation reports accessed through an Official Information Request or a comparison with reports in the international domain might also contribute to a more strongly warranted conclusion. Finally, in the event of evaluation becoming an accredited profession in Aotearoa New Zealand, action-based evaluative research which tracks changes in evaluative reasoning practices over time would provide further insights.

References

Abma, T. A. (2006). The social relations of evaluation. In I. F. Shaw, J. C. Greene & M. M. Mark (Eds.), *The Sage handbook of evaluation* (pp. 185–200). London: Sage Publications.

Arens, S. A. (2005). *A study of evaluative reasoning in evaluation studies judged "outstanding"*. Unpublished doctoral thesis, Indiana University.

Booth, W. C., Colomb, G. G., & Williams, J. M. (2008). *The craft of research* (3 ed.). Chicago: The University of Chicago Press. http://dx.doi.org/10.7208/chicago/9780226062648.001.0001

Chelimsky, E. (1997). The coming transformations in evaluation. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century* (pp. 1–26). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781483348896.n1

Chelimsky, E. (2012). Balancing evaluation theory and practice in the real world. *American Journal of Evaluation*, 34(1), 91–98. http://dx.doi.org/10.1177/1098214012461559

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Thousand Oaks, CA: Sage.

Denzin, N. K., & Lincoln, Y. S. (2011). The discipline and practice of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (Vol. 4, pp. 1–19). Thousand Oaks, CA: Sage.

Donaldson, S. I., & Lipsey, M. W. (2006). Roles for theory in contemporary evaluation practice: developing practical knowledge. In I. F. Shaw, J. C. Greene & M. M. Mark (Eds.), *The Sage handbook of evaluation.* Thousand Oaks, CA: Sage.

Fournier, D. (1995). Establishing evaluative conclusions: A distinction between general and working logic. *New Directions for Evaluation 68*, 15–32. http://dx.doi.org/10.1002/ev.1017

Fournier, D., & Smith, N. L. (1993). Clarifying the merits of argument in evaluation practice. *Evaluation and Program Planning*, *16*, 315–323. http://dx.doi.org/10.1016/0149-7189(93)90044-9

Gluckman, P. (2013). *The role of evidence in policy formation and implementation*. Auckland: Office of the Prime Minister's Science Advisory Committee. Retrieved from http://www.pmcsa.org.nz/publications/

Greene, J. C. (2005). Context. In S. Matheson (Ed.), *Encyclopedia of evaluation* (pp. 82–84). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412950558.n108

Greene, J. C. (2011). The construct(ion) of validity as argument. *New Directions for Evaluation*, 130, 81–91. http://dx.doi.org/10.1002/ev.367

Grob, G. F. (2003). A truly useful bat is one found in the hands of a slugger. *American Journal of Evaluation, 24*(4), 507–514. http://dx.doi.org/10.1177/109821400302400407

Hare, R. M. (1967). What is a value judgment? In P. W. Taylor (Ed.), *Problems of Moral Philosophy—An Introduction to Ethics* (3rd ed.). Belmont, CA: Wadsworth.

Henry, G. T. (2002). Choosing criteria to judge program success: A values inquiry. *Evaluation*, *8*(2), 182–202. http://dx.doi.org/10.1177/1358902002008002513

House, E. R. (1980). *Evaluating with validity*. Beverley Hills, CA: Sage.

House, E. R. (1995). Putting things together coherently: Logic and justice. *New Directions for Evaluation, 68*(33–48). http://dx.doi.org/10.1002/ev.1018

House, E. R., & Howe, K. R. (1999). *Values in evaluation and social research.* Thousand Oaks, CA: Sage.

Hurteau, M., Houle, S., & Mongiat, S. (2009). How legitimate and justified are judgments in program evaluation? *Evaluation,15*(3), 307–319. http://dx.doi.org/10.1177/1356389009105883

Julnes, G. (2012). Developing policies to support valuing in the public interest. *New Directions for Evaluation*, 133. http://dx.doi.org/10.1002/ev.20012

Nutley, S., Davies, H., & Walter, I. (2003). Evidence-based policy and practice: Lessons from the United Kingdom. *Social Policy Journal of New Zealand*, 20, 29–48.

Patton, M. Q. (2012). Contextual pragmatics of valuing. *New Directions for Evaluation*, 133, 97–108. http://dx.doi.org/10.1002/ev.20011

Prime Minister and Minister for the Cabinet Office. (1999). *Modernising government*. (Cm 4310). London: Author. Retrieved from https://www.wbginvestmentclimate.org/uploads/modgov.pdf

Rescher, N. (1969). *Introduction to value theory*. Englewood Cliffs, NJ: Prentice-Hall.

Schwandt, T. A. (1997). The landscape of values in evaluation: Charted terrain and unexplored territory. *New Directions for Evaluation*, 76, 25–38. http://dx.doi.org/10.1002/ev.1085

Schwandt, T. A. (2001). Responsiveness and everyday life. *New Directions for Evaluation*, 92, 73–86. http://dx.doi.org/10.1002/ev.36

Schwandt, T. A. (2008). Educating for intelligent belief in evaluation. *American Journal of Education*, *29*(2), 138–150. http://dx.doi.org/10.1177/1098214008316889

Scriven, M. (1980). *The logic of evaluation*. Thousand Oaks, CA: Edgepress.

Scriven, M. (1991). *Evaluation thesaurus* (4th edn). Newbury Park, CA: Sage.

Scriven, M. (1994). The final synthesis. *Evaluation Practice*, *15*(3), 367–382. http://dx.doi.org/10.1016/0886-1633(94)90031-0

Scriven, M. (1995). The logic of evaluation and evaluation practice. *New Directions for Evaluation*, *68*, 49–70. http://dx.doi.org/10.1002/ev.1019

Scriven, M. (2013). The foundations and future of evaluation. In S. I. Donaldson (Ed.), *The future of evaluation in society: A tribute to Michael Scriven* (pp. 11–44). Charlotte, NC: Information Age Publishing.

Scobie, G. (2009, August). *Evidence-based policy: reflections from New Zealand*. Paper presented at the Strengthening Evidence-based Policy Conference in the Australian Federation, Canberra. Retrieved from http://www.pc.gov.au/research/completed/strengthening-evidence

Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation,* 19(1), 1–19. http://dx.doi.org/10.1177/109821409801900102

Smith, N. J. (1995). The influence of societal games on the methodology of evaluative inquiry. *New Directions for Evaluation*, 68, 5–14. http://dx.doi.org/10.1002/ev.1016

Stake, R. E. (2004). *Standards-based and responsive evaluation*. Thousand Oaks, CA: Sage.

Stake, R. E., & Schwandt, T. A. (2006). On discerning quality in evaluation. In I. F. Shaw & J. C. Greene (Eds.), *The Sage Handbook of Evaluation*. London: Sage Publications.

Taylor, P. W. (1961). *Normative discourse*. Englewood Cliffs, NJ: Prentice-Hall

Toulmin, S., Rieke, R., & Janik, A. (1979). *An introduction to reasoning*. New York: Macmillan Publishing.

Valovirta, V. (2002). Evaluation utilisation as argumentation. *Evaluation*, *8*(1), 60–80. http://dx.doi.org/10.1177/1358902002008001487

The Authors

Heather Nunns, Analytic Matters.
Email: heather@analyticmatters.co.nz
Associate Professor Robin Peace, College of Humanities & Social Sciences,
Massey University.
Email: R.Peace@massey.ac.nz
Professor Karen Witten, Centre for Social and Health Outcomes Research and Evaluation, Massey University.
Email: K.Witten@massey.ac.nz